格助詞パターンの実態

荻野孝野†,植田禎子†,小林正博†,井佐原均††

†日本システムアプリケーション，　††独立行政法人 情報通信研究機構

1　はじめに

日本語文を解析するにあたって、述部にかかる「体言＋格助詞」と述部は、文を構成する重要な要素となる。これらの関係は一般に「結合価」と呼ばれている。これは、「(子供)に(本)を渡す」の「渡す」は、格助詞「～ニ～ヲ」が必要であるというように、動詞側が必要とする体言の意味グループや格助詞の関係に着目したもので、言語処理の分野では１９７０年代から行われ、様々な自然言語処理の土台として導入されてきたと言える。日本国内のこの分野の言語的研究としては 1970 年代初期から海外の結合価研究に着目した石綿敏雄氏（石綿　1975）によって、動詞の分析が試みられ、その後、IPAL 動詞辞書(IPA 1987)やNTT構文体系(池原他　1997)などで結合価に着目したデータ作成が行われてきた。本研究では、結合価を①格助詞の組み合わせと②体言の意味特徴という二つの段階的要素からとらえ、ここでは、そのうち、①格助詞の組み合わせのみに着目して把握した、動詞にかかる格助詞パターンの実態について述べる。

本研究の特徴は、大量の実在する日本語文から検討した格助詞パターンの実態にあるといえる。著者らは、大量のコーパスの中で、日本語の格助詞パターンがどのような分布で出現するか、また本来必要な格を伴った格助詞パターンと、実際のコーパスに出現する格助詞パターンとの関係はどのようになるかなどについて、EDR 共起辞書から作成した「日本語動詞の結合価」(荻野他　2003)の動詞約 12,400 概念を対象に、調査、検討した。本研究は(1)格助詞パターンの出現形の実態調査、(2)出現形と基本形の区分、(3)基本形パターンの異なりの抽出という手順で行われた。これらの分析によって、日本語表現において格助詞がどのように出現するかを把握し、動詞ごとの基本形を確定することによって一般的に動詞固有の基本的な格がいつでもすべて出現するとは限らない日本語文の解析に役立てることをめざしたものである。

2　格助詞パターン抽出の元データ

本調査の対象となったデータは、ＥＤＲコーパスから作成されたＥＤＲ共起辞書（日本電子化辞書研究所 1995）から約 12,400 概念の動詞にかかる格関係のデータを抽出して作成した「日本語動詞の結合価」データ(例１)である。

（1）「日本語動詞の結合価」データ

例１（付表部分）は、ＥＤＲコーパス、そしてＥＤＲ共起辞書を元に作成された「日本語動詞の結合価」データの部分で、動詞にかかる体言部分が所定の格助詞の列に収めた形式で整理されているものである。

（2）「日本語動詞の結合価」データにおける格助詞の扱い

格助詞部分は、左から、「は」、「が」、「を」、「に」、「へ」、「から」、「より」、「まで」、「で」、「と」、「その他」の順に列が定められている。「日本語動詞の結合価」データの格助詞部分は、元データとなったＥＤＲコーパスから係り受け関係を抽出したものをベースにして作成されたものであるが、結合価データとしてまとめる段階で、手作業にて実データを解釈し、必要な部分には「態の変換」や「係助詞の実質格への置き換え」などを施し、格助詞を実質的な列に配置している。

3　格助詞パターンの抽出

3.1　作業手順の概要

「日本語動詞の結合価」データを基データとして以下の作業を行う。

（1）出現形の抽出：動詞ごとに格助詞の組み合わせをすべて抽出し（表１：単語概念別格助詞パターン）、出現形パターンの異なりを把握する（表２：格助詞パターンの出現形別の動詞一覧、表３：出現形の異なりパターンの部分）

（2）基本形と派生形の区分け：抽出した出現形を基本形と派生形の関係に分ける（表1「基本格サインとグループ化」の列）。

（3）基本形の抽出：（2）で設定した基本形について、（1）と同様に基本形パターンの異なりを把握する（表４：基本形の異なりパターンの部分）。

3.2　動詞内の出現形のグルーピングと基本形設定

動詞の格助詞パターンは、いつも必要な格がすべて表現された基本形で出現するとは限らない。表現の状態によって基本的な格助詞（必須格）の一部が省略されたり、逆にどの動詞にもつきうるような格助詞（任意格）が付加されている場合がある。動詞12,416概念について、格助詞の省略や追加を考慮しながら、格助詞パターンの基本形パターンと派生形パターンの関連を把握し、格助詞パターンのグループ化を行い、かつその中の基本パターンを設定した。

3.3　格助詞パターンの数量的検討

動詞概念の事例約 155,000 について、格助詞の組み合わせを抽出し、異なりパターンを出した。元データから抽出された約 35,500 の延べパターンは、出現形レベルおよび基本形レベルで表５に示すような異なりパターンに集約された。

表5　出現形、基本形の延べパターン数と異なりパターン数

	調査した概念数	調査対象に含まれた事例数	延べパターン数	異なりパターン数
出現形	12,416	155,433	35,526	173
基本形			13,353	84

3.4　動詞が必要とする格助詞の分布

次にこれらの格助詞パターンについて、意味的に重要度の高い格助詞でとらえたら、格助詞の分布はどうなるかについて検討した。これは動詞を表現するのに必要な格助詞に「ヲ、ニ(ヘ)、ト、カラ（ヨリ)、マデ」のような優先順をつけて、その格助詞パターンに含まれる優先度の高い二つの格助詞の組み合わせに限定し、格助詞パターンを単純化し、格助詞の分布を把握したものである。表6は、格助詞パターンについて格助詞の優先順に第1分類､第2分類の種別を入れたものの一部である。

表6　格助詞パターンの簡略化

元の格助詞パターン	変換後の格助詞パターン	第1分類	第2分類
_WO_HE_DE	_WO_NI_DE	WO	NI
_WO_HE_KR	_WO_NI_KR	WO	NI
_WO_HE_KR_DE	_WO_NI_KR_DE	WO	NI
_GA_WO_HE	_WO_NI	WO	NI

図1は、表3に示した出現形分布の第1格助詞に着目して、その割合を図示したものである。図1の第1格助詞レベルの分類でみると、「物を買う」のように「対象格」を取るものが動詞の約55%、その他は、「雨が降る」のように「現象」を示す動詞や、「歩く、走る」のように対象格を必要としない「ガ」格だけの動詞が18%、「右に曲がる」のように格助詞の中に「ヲ」を含まず「ニ」を含む動詞が17%で、以上「ガ、ヲ、ニ」のいずれかを第一格にする動詞で全体の約90%を占めることなどがわかった。図2は基本形パターンからの第1格助詞の分布である。基本形なのでさらに「ヲ」格の割合が上がっている。

4　まとめ

著者らは、出現数にして延べ約155,000の動詞概念を含む文例から、格助詞出現形において延べで35,500のパターン、異なりで173のパターンを抽出し、さらにこれらの出現形パターンを基本形と格助詞の省略や追加による派生形パターンとの関係でとらえ、動詞ごとに基本形パターンを設定した。こうして作成された基本形パターンは、延べにして13,350パターン、異なりにして84パターンに集約された。また、これらの格助詞パターンにおいて格助詞がどのように分布しているかを把握した。以上、これらの格助詞パターンが、省略格の推定や参照関係の把握など格助詞パターンを手がかりにした自然言語処理の基礎データとして活用することができればと思う。

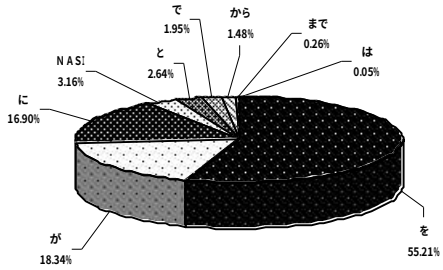

図1　出現形による格助詞の分布（第1格助詞）

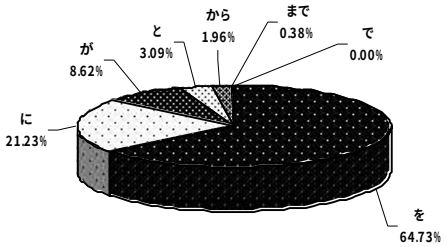

図2　基本形による格助詞の分布（第1格助詞）

参考文献

池原悟,宮崎正弘,白井諭,横尾昭男,中岩浩己,小倉健太郎,大山芳史,林良彦（1997）.日本語語彙大系.岩波書店

石綿敏雄（1975）.“日本語の生成語彙論的記述と言語処理の応用”.電子計算機による国語研究Ⅶ.国研報告54.秀英出版

荻野孝野,小林正博,井佐原均(2003).日本語動詞の結合価.三省堂

情報処理振興事業協会（1987）.計算機用日本語動詞辞書IPAL.情報処理振興事業協会(IPA)

日本電子化辞書研究所(1995).ＥＤＲ電子化辞書仕様説明書（第２版）EDR-TR045.日本電子化辞書研究所

独立行政法人情報通信研究機構.　EDR 電子化辞書.http://www2.nict.go.jp/kk/e416/EDR/J_index.html

例1　「日本語動詞の結合価データ」の部分

##### タベ･ル[食べる] 食べ/3bc6f0{食物をとる}						
受け側	:	は(58)	が(27)	を(141)	に(18)	例文
食べ			も (私 (名詞))	が(食べ物(名詞))		ところが実は私も、こちらの食べ物が食べられない。
食べ			は(日本人(名詞))	しか(シシャモ(名詞))		日本人は卵をもった子持ちシシャモしか、食べない。
食べ			は(わたし(名詞))	でも (何 (名詞))		わたしは、何でも食べられる。
食べ			は (私 (名詞))	が(シューマイ(名詞))		私はシューマイが無性に食べたくなった。

表 1 単語概念別格助詞パターン

# 単語 (概念別) ID	#単語見出し	# 単語 (概念別) ごとのパターン数	# 単語 (概念別) ごとの頻度	頻度	基本格サインとグループ化★	格助詞パターン	TOの細分類
1	ア･ウ[会う]	12	244	31	#1	_GA_NI	
1	ア･ウ[会う]	12	244	10	1	_GA_NI_DE	
1	ア･ウ[会う]	12	244	32	1	_NI_DE	
1	ア･ウ[会う]	12	244	48	1	_NI	
1	ア･ウ[会う]	12	244	25	#2	_GA_TO	/PAR
1	ア･ウ[会う]	12	244	11	2	_DE_TO	/PAR
1	ア･ウ[会う]	12	244	15	2	_GA_DE_TO	/PAR
1	ア･ウ[会う]	12	244	31	2	_TO	/PAR
1	ア･ウ[会う]	12	244	12	12	_DE	
1	ア･ウ[会う]	12	244	7	12	_GA_DE	
1	ア･ウ[会う]	12	244	11	12	_GA	
1	ア･ウ[会う]	12	244	11	12	格助詞ナシ	

表2　格助詞パターンの出現形別の動詞一覧の部分

格助詞パターン	概念数	出現頻度合計	動詞一覧 ()内は出現頻度
_WO	4443	32389	ア・ム[編む](3),ア・ム[編む](2),アイ・スル[愛する](10),アイテド・ル[相手取る](1),アイドク・スル[愛読する](1),アイヨウ・スル[愛用する](1),アオ・グ[仰ぐ](6) ,アオギミ・ル[仰ぎ見る](2),アカ・ス[明かす](8),アガナ・ウ[あがなう](1),アキナ・ウ[商う](1),アキラメ・ル[諦める](2), ,アクセス・スル[アクセスする](2),;
_GA	4038	22240	ア・ウ[会う](11),ア・ウ[合う](24),ア・ク[開く](6),ア・ク[空く](9),ア・ム[編む](1),ア・ル[ある](347),アイ・スル[愛する](4),アイコウ・スル[愛好する](1),;

表3　出現形の異なりパターンの部分

	格パターン	概念異なり数	概念の割合	概念出現頻度	出現割合	第1分類	第2分類
1	_WO	4443	12.8%	32389	20.8%	WO	
2	_GA_WO	3806	11.0%	24207	15.6%	WO	
3	_WO_DE	2057	5.9%	8001	5.1%	WO	
4	_GA_WO_DE	1226	3.5%	3227	2.1%	WO	
:							
9	_WO_NI	1648	4.7%	8662	5.6%	WO	NI
10	_GA_WO_NI	1004	2.9%	3254	2.1%	WO	NI
11	_WO_NI_DE	375	1.1%	667	0.4%	WO	NI
:							

表4　基本形の異なりパターンの部分

	格パターン	概念異なり数	概念の割合	概念出現頻度	出現割合	第1分類	第2分類
1	_GA_WO	4393	32.90%	43552	23.85%	WO	
2	_GA_WO_DE	577	4.32%	14084	7.71%	WO	
3	_WO_DE	69	0.52%	1372	0.75%	WO	
4	_HA_GA_WO	31	0.23%	85	0.05%	WO	
:							
9	_GA_WO_NI	1730	12.96%	32562	17.83%	WO	NI
10	_GA_WO_NI_DE	64	0.48%	4006	2.19%	WO	NI
11	_GA_WO_HE	105	0.79%	2342	1.28%	WO	NI
12	_GA_WO_NI_KR	59	0.44%	1986	1.09%	WO	NI
13	_GA_WO_HE_KR	28	0.21%	826	0.45%	WO	NI